Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 11 January 2013 (MN LATEX style file v2.2)

On statistical uncertainty in nested sampling

Charles R. Keeton
Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854 USA

11 January 2013

arXiv:1102.0996v1 [astro-ph.IM] 4 Feb 2011

ABSTRACT
Nested sampling has emerged as a valuable tool for Bayesian analysis, in particular for determining the Bayesian evidence. The method is based on a specific type of random sampling of the likelihood function and prior volume of the parameter space. I study the statistical uncertainty in the evidence computed with nested sampling. I examine the uncertainty estimator from Skilling (2004, 2006) and introduce a new estimator based on a detailed analysis of the statistical properties of nested sampling. Both perform well in test cases and make it possible to obtain the statistical uncertainty in the evidence with no additional computational cost.

1 INTRODUCTION

Bayesian statistics provide a general framework for confronting models with data (e.g., Gelman et al. 1995). Constraints on model parameters are quantified by the *posterior distribution* for the parameters given the data. The overall quality of a model is characterised by an integral over the posterior, which is known as the *evidence*. The Bayesian evidence is especially valuable as an objective means of comparing models with different numbers of parameters.

The challenge with Bayesian statistics is finding an efficient method to explore the posterior and/or compute the evidence. The posterior may occupy many dimensions and have a complicated (and possibly multi-modal) shape. Markov Chain Monte Carlo (MCMC) methods have become popular as a way to generate samples of points drawn from arbitrary posteriors (e.g., Gelman et al. 1995). MCMC samples are great for inferring parameter values and ranges, but they cannot be used by themselves to evaluate the evidence. MCMC methods can be extended to yield the evidence via thermodynamic integration (see Gelman & Meng 1998, and references therein), but that approach can be computationally intensive.

Skilling (2004, 2006) recently introduced an approach called nested sampling that is specifically designed to compute the Bayesian evidence. Roughly speaking, the idea is to peel away layers of constant likelihood one by one, and combine the likelihood values with the volumes of the layers to obtain the evidence. The volumes may be difficult to determine, but they can be estimated statistically if the likelihood layers are chosen in a particular way (see § 2.1 for details). While the analysis focuses on the evidence, it can yield a set of points drawn from the posterior as a natural by-product.

There are two practical challenges with nested sampling. The first is that at each step we need to generate a new point drawn from the region inside an iso-likelihood surface. A lot of the literature on nested sampling addresses methods for picking new points. Mukherjee et al. (2006) discuss drawing points inside a multi-dimensional ellipsoid that encloses the likelihood surface at $\mathcal{L}_0$, and ignoring any that have $\mathcal{L} < \mathcal{L}_0$. Shaw et al. (2007), Feroz & Hobson (2008), and Feroz et al. (2009) develop methods that use multiple ellipsoids to handle more complicated likelihood functions, including multi-modal distributions. Chopin & Robert (2008) point out that importance sampling can be powerful if one can find a distribution that is easy to draw from and approximates the likelihood distribution moderately well. Betancourt (2010) advocates using constrained Hamiltonian Monte Carlo methods to evolve a new point from one of the known points. All of those methods keep the core approach of peeling away likelihood layers in sequence from the outside in, and differ only in the details of picking new points. Brewer et al. (2009) introduce a variant they call diffusive nested sampling that does not always require the steps to proceed from the outside in.

The second challenge is that nested sampling, like any stochastic sampling procedure, has some statistical uncertainty in its results. General proofs establish that nested sampling converges to the correct evidence with an error that scales as $N^{-1/2}$ where N is a measure of the computational effort (Chopin & Robert 2008; Skilling 2009). However, in practical applications it would be nice to have a specific estimate of the statistical uncertainty in the evidence. That is the purpose of this paper. I first review the nested sampling procedure (§ 2.1) and a popular estimator from Skilling (2004, 2006) for the statistical uncertainty in the evidence (§ 2.2). I then introduce a new uncertainty estimator based on an analysis of the statistical

© 0000 RAS

properties of the nested sampling procedure (§§ 2.3 and 2.4). I use numerical tests to assess the estimators and provide some guidelines for choosing parameters that control the number of samplings (§ 3). The results presented here are applicable to any implementation of nested sampling that uses the conventional approach of peeling away likelihood layers in one direction only (i.e., to all current methods other than diffusive nested sampling).

2 THEORETICAL FRAMEWORK

2.1 Nested sampling

To establish the concepts and notation, it is useful to review the nested sampling algorithm (see Skilling 2004, 2006 for details). Consider a likelihood function $\mathcal{L}(\theta)$ defined on a parameter space θ, which may be multi-dimensional.[1] Priors on the parameters are specified by $\pi(\theta)$, which is normalised such that $\int \pi(\theta)\ d\theta = 1$. With simple flat priors, $\pi(\theta) = 1/V$ where V is the volume spanned by the allowed range of parameters, but the framework can incorporate non-flat priors as well. The Bayesian evidence is then

$$Z = \int \mathcal{L}(\theta)\ \pi(\theta)\ d\theta \tag{1}$$

Define a function $X(L)$ to be the fraction of the prior volume that lies at a likelihood level higher than L:

$$X(L) = \int_{\mathcal{L}(\theta)>L} \pi(\theta)\ d\theta \tag{2}$$

This is a monotonic decreasing function, with $X(0) = 1$. In principle, we can invert to find $L(X)$ and then rewrite eq. (1) as

$$Z = \int_0^1 L(X)\ dX \tag{3}$$

Now suppose we can generate a sample of $N_{\rm nest}$ points $\{L_i, X_i\}$ such that the likelihood increases while the fractional volume decreases with the index i; in other words, $L_i > L_{i-1}$ and $X_i < X_{i-1}$, and we can consider $L_0 = 0$ and $X_0 = 1$. Then we can discretise the integral to estimate the evidence as

$$Z = \sum_{i=1}^{N_{\rm nest}} L_i (X_{i-1} - X_i) \tag{4}$$

Later it will be useful to consider the buildup of evidence by examining the "partial evidence" due to the contribution from the first k steps:

$$Z_k = \sum_{i=1}^{k} L_i (X_{i-1} - X_i) \tag{5}$$

There is some error in eq. (4) associated with approximating the integral as a sum, but it is generally small compared with the statistical uncertainty (Skilling 2006). There is also some error induced by truncating the sum, which is discussed in § 2.4.

The heart of nested sampling is the method for generating the likelihood sampling $\{L_i\}$ and volume sampling $\{X_i\}$. The idea is that it is (relatively) straightforward to produce a relevant likelihood sampling, but it can be difficult to determine the associated volumes so we treat those statistically. Consider some likelihood threshold $\mathcal{L}_0$ enclosing a volume $\mathcal{V}_0$. Suppose we have M points drawn *uniformly* from that volume. In general there will be some (slightly) higher likelihood threshold $\mathcal{L}_1 > \mathcal{L}_0$ that encloses all M points. Statistically speaking, we can estimate the smaller enclosed volume as $\mathcal{V}_1 = \mathcal{V}_0 t_1$ where t_1 is the largest of M random numbers drawn uniformly between 0 and 1. In other words, t_1 is drawn from the probability distribution for the largest of M uniform deviates between 0 and 1, which is

$$p(t) = M\, t^{M-1} \quad \text{for } t \in [0,1] \tag{6}$$

We can generalise to non-uniform priors by defining the "volumes" to be integrals of the priors over the relevant regions and having the M points drawn from the prior distribution. The probability distribution for t_1 remains unchanged.

That idea leads to the following procedure. Begin with M points—known as "live" points—drawn uniformly from the full prior distribution. Let the likelihoods of the live points be $\mathcal{L}_\mu$ for $\mu = 1, \ldots, M$. Then at step k of the nested sampling:

(i) Extract the lowest likelihood live point and call it the k-th sampled point: $L_k = \min(\mathcal{L}_\mu)$.

(ii) Estimate the associated volume as

$$X_k = X_{k-1} t_k \tag{7}$$

where t_k is a random number drawn from $p(t)$ in eq. (6).

(iii) Replace the extracted live point with a new point that is drawn from the priors but restricted to the region $\mathcal{L}(\theta) \geqslant L_k$.

[1] To simplify the notation, I do not explicitly indicate vectors or write the data dependence in the likelihood function.

Iterating this process for a total of $N_{\rm nest}$ steps yields a likelihood sampling $\{L_i\}$ and volume sampling $\{X_i\}$ that can be combined using eq. (4) to estimate the evidence. This is the conventional nested sampling technique as defined by Skilling (2004, 2006). The various implementations of nested sampling mainly differ in the way they find the replacement point in step (iii).

2.2 Skilling's error analysis

To estimate the statistical uncertainty associated with stochastic sampling, Skilling (2004, 2006) invokes information theory. In general the posterior $p(\theta)$ is (much) narrower than the prior $\pi(\theta)$, and we can characterise the difference in terms of the "information gain" (also known as the Kullback-Leibler divergence; see Kullback 1959)

$$H = \int p(\theta) \ln \frac{p(\theta)}{\pi(\theta)} \, d\theta \tag{8}$$

By Bayes's theorem, $p(\theta) = \mathcal{L}(\theta)\pi(\theta)/Z$ so we can write

$$H = \frac{1}{Z} \int \mathcal{L}(\theta)\pi(\theta) \ln \frac{\mathcal{L}(\theta)}{Z} \, d\theta = \frac{1}{Z} \int L(X) \ln L(X) \, dX \; - \; \ln Z \tag{9}$$

using the same change of variables as in eq. (3). This integral can be discretised just like the evidence integral, so it is straightforward to estimate H from a given sampling $\{L_i, X_i\}$.

Skilling (2004, 2006) argues that the number of steps needed to reach the posterior is approximately HM where M is the number of live points, and that the dominant statistical uncertainty arises from Poisson fluctuations $\sqrt{HM}$ in the number of steps. Thus, he estimates an uncertainty in $\ln Z$ of about $\sqrt{H/M}$. Note that Skilling argues that $\ln Z$, and not Z itself, is the quantity likely to have a fairly symmetric and quasi-Gaussian distribution. However, if the uncertainty is small (specifically, $\sigma_Z/Z \ll 1$), then Z itself will also be Gaussian distributed and Skilling's estimate corresponds to a fractional uncertainty in the evidence of

$$\frac{\sigma_Z}{Z} \approx \sqrt{\frac{H}{M}} \tag{10}$$

This estimator is often used in nested sampling applications, but its accuracy has not (to my knowledge) been rigorously established.

2.3 Moment-based error analysis

Skilling (2006) mentions that it should be possible to obtain a more detailed estimate of the statistical uncertainty by computing the mean and variance of Z over all possible realisations of the volume sampling $\{X_i\}$, but he does not carry out the analysis. The goal of this section is to compute $\langle Z \rangle$ and $\langle Z^2 \rangle$ to obtain a new estimator for σ_Z. Since this estimator is based on the standard deviation, it is most useful when Z is Gaussian distributed, i.e., when the uncertainties are small ($\sigma_Z/Z \ll 1$). This does not seem like a significant limitation, though, because in many applications it will be desirable to achieve small uncertainties.

It is convenient to use eq. (7) to write the volumes as

$$X_i = \prod_{j=1}^{i} t_j \tag{11}$$

The advantage is that the X_i's are statistically correlated, but the t_i's are independent and that allows us to decompose the joint probability density for all the t_i's into a product:

$$p_{\rm all}(t_1, t_2, t_3, \ldots) = p(t_1) \; p(t_2) \; p(t_3) \cdots \tag{12}$$

where $p(t)$ is from eq. (6). We can then write the average of any quantity f over all realisations of the volume sampling as

$$\langle f \rangle \equiv \int f(t_1, t_2, t_3, \ldots) \; p(t_1) \; p(t_2) \; p(t_3) \cdots \; dt_1 \; dt_2 \; dt_3 \cdots \tag{13}$$

It is important to understand that such an average only spans the volume sampling; at this point we are not considering different realisations of the likelihood sampling. As part of this analysis we need moments of the t probability distribution,

$$\langle t^n \rangle = \int_0^1 t^n \; p(t) \; dt = \frac{M}{M+n} \tag{14}$$

Combining eqs. (4) and (11), we can write the (partial) evidence in terms of the t_i's as

$$Z_k = \sum_{i=1}^{k} L_i (1 - t_i) \prod_{j=1}^{i-1} t_j = \sum_{i=1}^{k} L_i \left(\prod_{j=1}^{i-1} t_j - \prod_{j=1}^{i} t_j \right) \tag{15}$$

Since the terms in the products are statistically independent, we can factorise the average of a product and write

$$\left\langle \prod_{j=1}^{n} t_j \right\rangle = \prod_{j=1}^{n} \langle t_j \rangle = \langle t \rangle^n \tag{16}$$

for any $n \in [1, N_{\rm nest}]$. This allows us to write the average of the evidence after any step k as

$$\langle Z_k \rangle = \left\langle \sum_{i=1}^{k} L_i \left(\prod_{j=1}^{i-1} t_j - \prod_{j=1}^{i} t_j \right) \right\rangle = \sum_{i=1}^{k} L_i \left(\langle t \rangle^{i-1} - \langle t \rangle^{i} \right) = \frac{1}{M} \sum_{i=1}^{k} L_i \langle t \rangle^i = \frac{1}{M} \sum_{i=1}^{k} L_i \left(\frac{M}{M+1} \right)^i \tag{17}$$

This is a simple expression for the (partial) evidence averaged over all possible realisations of the volume sampling (given a particular likelihood sampling $\{L_i\}$). Obviously the final evidence is obtained just by evaluating at $k = N_{\rm nest}$.

To compute the second moment it is convenient to begin with the partial evidence from eq. (5):

$$\begin{aligned} \left\langle Z_k^2 \right\rangle &= \left\langle \left[\sum_{i=1}^{k} L_i (1 - t_i) \prod_{j=1}^{i-1} t_j \right] \left[\sum_{i'=1}^{k} L_{i'} (1 - t_{i'}) \prod_{j'=1}^{i'-1} t_{j'} \right] \right\rangle \\ &= \left\langle Z_{k-1}^2 \right\rangle + \left\langle \left[L_k (1 - t_k) \prod_{j=1}^{k-1} t_j \right]^2 \right\rangle + 2 \left\langle \left[\sum_{i=1}^{k-1} L_i (1 - t_i) \prod_{j=1}^{i-1} t_j \right] \left[L_k (1 - t_k) \prod_{j'=1}^{k-1} t_{j'} \right] \right\rangle \end{aligned} \tag{18}$$

In the second line I separate the joint sum over $i, i' \leqslant k$ into three components. The first component includes all terms with $i, i' \leqslant k-1$, so we can immediately recognise it as $\left\langle Z_{k-1}^2 \right\rangle$. The second component is the term with $i = i' = k$. The third component includes all terms in which one index equals k while the other runs over values $\leqslant k-1$. Since we can interchange i and i', there is a leading factor of 2.

It takes a few steps to evaluate the averages. First consider the second term in eq. (18). Writing out the products of t_i's and collecting terms yields

$$\left\langle L_k^2 \left(\prod_{j=1}^{k-1} t_j^2 - 2 t_k \prod_{j=1}^{k-1} t_j^2 + \prod_{j=1}^{k} t_j^2 \right) \right\rangle = L_k^2 \left(\left\langle t^2 \right\rangle^{k-1} - 2 \langle t \rangle \left\langle t^2 \right\rangle^{k-1} + \left\langle t^2 \right\rangle^{k} \right) = \frac{2}{M(M+1)} L_k^2 \left\langle t^2 \right\rangle^k \tag{19}$$

Now consider the third term in eq. (18). We can rewrite the products, taking care to distinguish the t's that appear twice in a product from those that appear just once, and thus obtain

$$\begin{aligned} &\left\langle 2 L_k \sum_{i=1}^{k-1} L_i \left(\prod_{j=1}^{i-1} t_j^2 \prod_{j'=i}^{k-1} t_{j'} - \prod_{j=1}^{i} t_j^2 \prod_{j'=i+1}^{k-1} t_{j'} - \prod_{j=1}^{i-1} t_j^2 \prod_{j'=i}^{k} t_{j'} + \prod_{j=1}^{i} t_j^2 \prod_{j'=i+1}^{k} t_{j'} \right) \right\rangle \\ &= 2 L_k \sum_{i=1}^{k-1} L_i \left(\left\langle t^2 \right\rangle^{i-1} \langle t \rangle^{k-i} - \left\langle t^2 \right\rangle^{i} \langle t \rangle^{k-i-1} - \left\langle t^2 \right\rangle^{i-1} \langle t \rangle^{k-i+1} + \left\langle t^2 \right\rangle^{i} \langle t \rangle^{k-i} \right) \\ &= \frac{2}{M(M+1)} L_k \langle t \rangle^k \sum_{i=1}^{k-1} L_i \frac{\left\langle t^2 \right\rangle^i}{\langle t \rangle^i} \end{aligned} \tag{20}$$

Notice that eq. (19) has the same form as each term in the sum in eq. (20), but with index $i = k$. So when we insert eqs. (19) and (20) back into eq. (18), we can write

$$\left\langle Z_k^2 \right\rangle = \left\langle Z_{k-1}^2 \right\rangle + \frac{2}{M(M+1)} L_k \langle t \rangle^k \sum_{i=1}^{k} L_i \frac{\left\langle t^2 \right\rangle^i}{\langle t \rangle^i} \tag{21}$$

With this expression for the second moment of the partial evidence, we see that the second moment of the full evidence can be written as

$$\left\langle Z^2 \right\rangle = \frac{2}{M(M+1)} \sum_{k=1}^{N_{\rm nest}} L_k \langle t \rangle^k \sum_{i=1}^{k} L_i \frac{\left\langle t^2 \right\rangle^i}{\langle t \rangle^i} = \frac{2}{M(M+1)} \sum_{k=1}^{N_{\rm nest}} L_k \left(\frac{M}{M+1} \right)^k \sum_{i=1}^{k} L_i \left(\frac{M+1}{M+2} \right)^i \tag{22}$$

Combining eqs. (17) and (22) in the usual way yields a new estimator for the statistical uncertainty in the evidence:

$$\sigma_Z^2 = \frac{2}{M(M+1)} \sum_{k=1}^{N_{\rm nest}} L_k \left(\frac{M}{M+1} \right)^k \sum_{i=1}^{k} L_i \left(\frac{M+1}{M+2} \right)^i - \frac{1}{M^2} \left[\sum_{i=1}^{N_{\rm nest}} L_i \left(\frac{M}{M+1} \right)^i \right]^2 \tag{23}$$

For comparison, rewriting eq. (10) in the current notation yields the following expression for Skilling's uncertainty estimator:

$$\sigma_Z^2 = \frac{1}{M^3} \left[\sum_{i=1}^{N_{\rm nest}} L_i \left(\frac{M}{M+1} \right)^i \right] \left[\sum_{j=1}^{N_{\rm nest}} L_j \ln L_j \left(\frac{M}{M+1} \right)^j \right] - \frac{1}{M^3} \left[\sum_{i=1}^{N_{\rm nest}} L_i \left(\frac{M}{M+1} \right)^i \right]^2 \ln \left[\frac{1}{M} \sum_{j=1}^{N_{\rm nest}} L_j \left(\frac{M}{M+1} \right)^j \right] \tag{24}$$

On the surface these two expressions look quite different, so it is interesting to compare them in quantitative examples.

2.4 Handling the remainder

When the nested sampling procedure is complete, there is some (small) remaining volume, $X_{N_{\rm nest}}$, whose contribution to the evidence is neglected in eq. (4). While we can make its contribution arbitrarily small by taking enough steps (see § 3.4), we can also include it at the expense of making the analytic expressions slightly more complicated.

Suppose we truncate nested sampling after step k and compute the partial "nested evidence" Z_k from eq. (5). We can estimate the remaining evidence as a product of the remaining volume, X_k, and the mean likelihood within that volume. Since the live points are drawn uniformly from X_k, we can estimate the mean likelihood from the live points as

$$\bar{\mathcal{L}}^{(k)} = \frac{1}{M}\sum_{\mu=1}^{M}\mathcal{L}_\mu \tag{25}$$

Here the overbar distinguishes this average over live points from an average over volume realisations, and the superscript is a reminder that the average is taken after step k. Thus the "live evidence" is

$$Z_k^{\rm live} = \bar{\mathcal{L}}^{(k)} X_k \tag{26}$$

Averaging the live evidence over all volume realisations yields

$$\left\langle Z_k^{\rm live}\right\rangle = \bar{\mathcal{L}}^{(k)}\left\langle X_k\right\rangle = \bar{\mathcal{L}}^{(k)}\left\langle\prod_{j=1}^{k} t_j\right\rangle = \bar{\mathcal{L}}^{(k)}\left(\frac{M}{M+1}\right)^k \tag{27}$$

The second moment is

$$\left\langle (Z_k^{\rm live})^2\right\rangle = (\bar{\mathcal{L}}^{(k)})^2\left\langle X_k^2\right\rangle = (\bar{\mathcal{L}}^{(k)})^2\left\langle\prod_{j=1}^{k} t_j^2\right\rangle = (\bar{\mathcal{L}}^{(k)})^2\left(\frac{M}{M+2}\right)^k \tag{28}$$

so the statistical uncertainty in the live evidence is

$$\sigma^2_{Z_k^{\rm live}} = (\bar{\mathcal{L}}^{(k)})^2\left(\frac{M}{M+1}\right)^k\left[\left(\frac{M+1}{M+2}\right)^k - \left(\frac{M}{M+1}\right)^k\right] \tag{29}$$

Now consider the estimate of the total evidence after step k,

$$Z_k^{\rm tot} = Z_k + Z_k^{\rm live} \tag{30}$$

The average over volume realisations is simply obtained from eqs. (17) and (27). The statistical uncertainty in $Z_k^{\rm tot}$ is

$$\sigma^2_{Z_k^{\rm tot}} = \sigma^2_{Z_k} + \sigma^2_{Z_k^{\rm live}} + 2\left(\left\langle Z_k Z_k^{\rm live}\right\rangle - \left\langle Z_k\right\rangle\left\langle Z_k^{\rm live}\right\rangle\right) \tag{31}$$

The term in parentheses accounts for the fact that $Z_k^{\rm live}$ and Z_k are not independent because they both involve the same volume sampling. The cross term can be evaluated using an analysis similar to that in eq. (20), which yields

$$\left\langle Z_k Z_k^{\rm live}\right\rangle = \frac{\bar{\mathcal{L}}^{(k)}}{M+1}\left(\frac{M}{M+1}\right)^k\sum_{i=1}^{k} L_i\left(\frac{M+1}{M+2}\right)^i \tag{32}$$

Putting the pieces together, we find that including the live evidence increases the statistical uncertainty in the total evidence according to

$$\sigma^2_{Z_k^{\rm tot}} = \sigma^2_{Z_k} + \Delta\sigma^2_{Z_k} \tag{33}$$

where the original uncertainty is given in eq. (23) while the increase is

$$\Delta\sigma^2_{Z_k} = (\bar{\mathcal{L}}^{(k)})^2\left(\frac{M}{M+1}\right)^k\left[\left(\frac{M+1}{M+2}\right)^k - \left(\frac{M}{M+1}\right)^k\right] + 2\bar{\mathcal{L}}^{(k)}\left(\frac{M}{M+1}\right)^k\sum_{i=1}^{k} L_i\left[\frac{1}{M+1}\left(\frac{M+1}{M+2}\right)^i - \frac{1}{M}\left(\frac{M}{M+1}\right)^i\right] \tag{34}$$

In the examples that follow, I take enough steps that the live evidence provides a negligible contribution by the end, but the formalism in this section can be used if the number of nested sampling steps is more modest.

3 NUMERICAL RESULTS

In this section I present numerical tests designed to assess the uncertainty estimators, and to investigate how many samples to use. Since the nested sampling framework does not require any specific assumptions about the form of the likelihood distribution, a Gaussian test case should be sufficient. However, I also consider a log-normal distribution as a check.

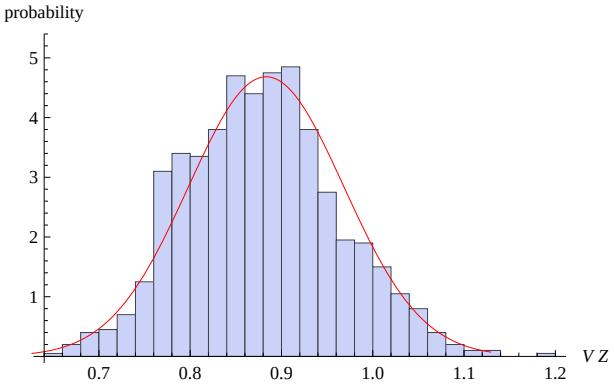

Figure 1. The histogram shows the distribution of evidence values (specifically VZ) for 1000 realisations of the volume sampling, given a particular likelihood sampling. The red curve shows a Gaussian distribution whose mean and varaiance are computed from eqs. (17) and (23). The mean and standard deviation from the simulations are 0.878 ± 0.084 (the mean differs from $VZ \approx 1$ for this particular likelihood sampling). For comparison, the analytic average over volume realisations is 0.883, and Skilling's and my uncertainty estimators yield 0.083 and 0.085, respectively.

3.1 Gaussian test case

Consider a multivariate Gaussian likelihood specified by some mean vector and covariance matrix. With flat priors we can make the following simplifications. Choose coordinates centred on the mean and aligned with the principal axes of the covariance matrix. Scale each coordinate by the standard deviation in that direction. This yields a multivariate Gaussian in canonical form,

$$\mathcal{L}(\theta) = (2\pi)^{-d/2}\ e^{-|\theta|^2/2} \tag{35}$$

where d is the number of dimensions. Let the prior volume be a cube of side length s centred on the origin, so the prior volume is $V = s^d$ and the priors are $\pi(\theta) = 1/V$. Thus the evidence is

$$Z = V^{-1} \int_V (2\pi)^{-d/2}\ e^{-|\theta|^2/2}\ d\theta \tag{36}$$

If the prior box is large enough to encompass essentially all of the likelihood, then $VZ \approx 1$ independent of the box size. For this reason, in the following tests I examine VZ instead of just Z. The information gain for this case is

$$H = \frac{1}{VZ} \int \mathcal{L} \ln \frac{\mathcal{L}}{Z}\ d\theta \approx -\frac{d}{2}\,(1 + \ln 2\pi) - \ln Z \tag{37}$$

In the last step I again assume the prior box is large.

For concreteness, I use a box with side length $s = 10$ in $d = 4$ dimensions; these choices influence the quantitative details but do not affect the general conclusions. In the fiducial case I use $M = 400$ live points and take $N_{\rm nest} = 4100$ steps (see § 3.4). The associated information gain is $H = 3.53$, and Skilling's estimator of the fractional uncertainty in the evidence has an analytic value of $\sqrt{H/M} = 0.094$.

3.2 Testing the volume sampling

I first generate a single realisation of the likelihood sampling and combine it with 1000 realisations of the volume sampling. Figure 1 shows a histogram of the VZ values from these direct simulations. The mean and standard deviation of the simulated values are 0.878 ± 0.084. The mean differs from the theoretical value $VZ \approx 1$ by about 1.5σ for this particular realisation of the likelihood sampling.

From eq. (17) the predicted average over volume realisations is 0.883. Skilling's estimator yields a statistical uncertainty of 0.083, while mine yields 0.085. The predicted Gaussian distribution agrees well with the simulation results, indicating that Z has a (nearly) Gaussian distribution when the uncertainties are small (qv. §§ 2.2 and 2.3). I conclude that the analytic expressions accurately describe the distribution of evidence values for many realisations of the volume sampling. It is striking that the two uncertainty estimators yield very similar values despite having such different analytic forms.

3.3 Testing the likelihood sampling

It is useful to see how the results vary with different realisations of the likelihood sampling. I now generate 1000 random likelihood samplings; for each one I compute the mean evidence averaged over all volume samplings using eq. (17). Figure 2 shows a histogram of the values of $\langle VZ \rangle_t$ for the different likelihood realisations (I add the subscript t to emphasise that the

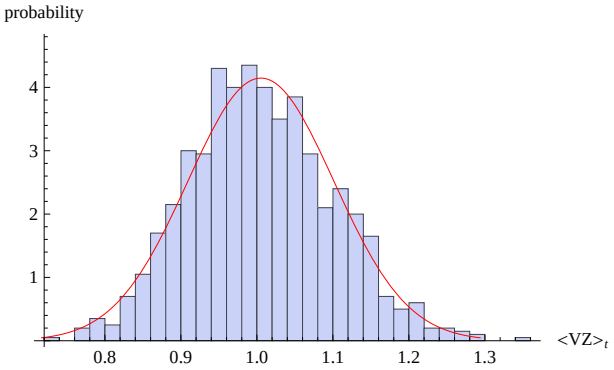

Figure 2. The histogram shows the distribution of $\langle VZ\rangle_t$ for 1000 realisations of the likelihood sampling. The notation $\langle VZ\rangle_t$ emphasises that I average over all volume samplings for each likelihood sampling. The red curve shows a Gaussian distribution whose mean and variance are computed from eqs. (17) and (23). The mean over all likelihood samplings is 1.005; the empirical scatter in the histogram is 0.094, while the predicted value is 0.094 for Skilling's estimator and 0.096 for mine.

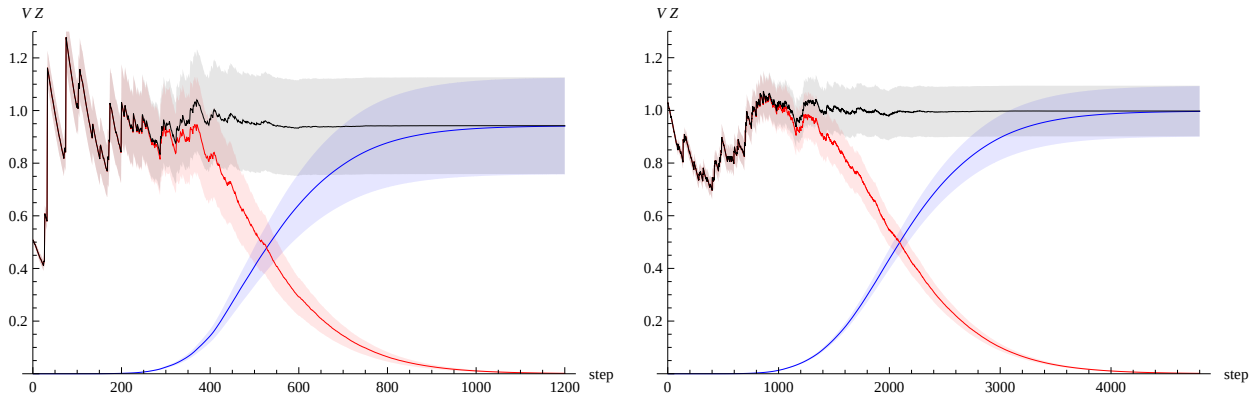

Figure 3. Development of the evidence as a function of step index. The blue band shows the mean and 1σ errors for the "nested evidence" (eqs. 17 and 23), the red band shows the "live evidence" (eqs. 27 and 29), and the black curve shows the total (eqs. 30 and 33). The number of live points is $M = 100$ (left) and $M = 400$ (right). With more live points, it takes more steps to reach convergence, but the ultimate uncertainty is smaller.

average is over volume samplings). The mean and standard deviation for the histogram are 1.005 ± 0.094. On average, nested sampling recovers the evidence very well.

Strictly speaking, both of the uncertainty estimators depend on the likelihood sampling, but the scatter across the likelihood realisations is $< 9\%$ so any single case provides a useful value. The average predicted uncertainty is 0.094 for Skilling's estimator, and 0.096 for mine. Also, the predicted Gaussian distribution agrees well with the empirical histogram. I conclude that the both analytic estimators characterise the statistical uncertainty in the evidence quite well. It is not obvious at this point why the two uncertainty estimators yield such similar results.

3.4 How many live points and steps?

Let us now consider how to choose the number of live points, M, and the number of nested sampling steps, $N_{\rm nest}$. One general goal is to have nested sampling "find" all significant modes in the posterior. The sampling procedure is basically guaranteed to find the peak for a unimodal distribution, but if the live points are too sparse they may miss some peaks (especially small ones) in a multi-model distribution. In order to have a reasonable probability of getting at least one live point in each mode at the outset, Feroz & Hobson (2008) suggest that the number of live points should exceed $V_{\rm prior}/V_{\rm min}$, where $V_{\rm prior}$ is the volume spanned by the priors while $V_{\rm min}$ is the volume of the smallest mode (which must be estimated since it cannot actually be known before the analysis is done).

The second consideration relates to achieving a robust and precise estimate of the evidence. Figure 3 shows the development of the evidence as a function of the step index, for two choices of M. After some number of steps the evidence and uncertainty saturate in the sense that taking additional steps does not significantly change the results. For a heuristic understanding, note that as nested sampling homes in on a likelihood peak the likelihood values become constant ($L_i \to L_{\rm peak}$) while the volumes become progressively smaller ($X_i \to 0$). For a rigorous proof of convergence, see Skilling (2009).

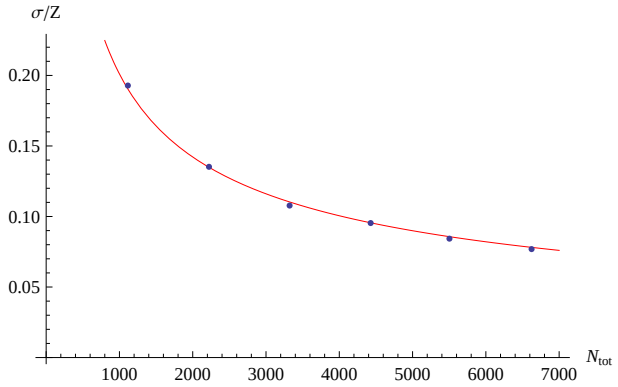

Figure 4. The points show the fractional uncertainty in the evidence versus the total number of likelihood evaluations ($N_{\rm tot} = N_{\rm nest} + M$) for tests in which the number of live points is $M = 100, 200, 300, 400, 500, 600$ (left to right). Here $N_{\rm nest}$ is determined for each M using the stopping threshold $\epsilon = 0.01$. The curve shows the scaling relation $\sigma_Z/Z \propto N_{\rm tot}^{-1/2}$.

The question arises of how to identify the point of diminishing returns. One simple possibility (Skilling 2004, 2006) is to compare the evidence accumulated through nested sampling (Z_k from eq. 5) with the remaining evidence estimated from the live points ($Z_k^{\rm live}$ from eq. 26). Figure 3 shows $Z_k^{\rm live}$ versus k in red. At early stages the curve shows a series of spikes: it rises sharply when a new live point is found that (temporarily) dominates the average over live points, then declines as the volume decreases. The curve smoothes out as the live points come to have more similar likelihoods, and then decays as the likelihoods saturate while the remaining volume continues to decrease. Roughly speaking, $Z^{\rm live}$ represents the evidence that has been "missed" by the nested sampling procedure, so we may want to continue the nested sampling until the ratio of live to nested evidence falls below some threshold: $Z^{\rm live}/Z < \epsilon$.

Figure 3 illustrates that using more live points means it takes more steps to reach a given ϵ threshold, but the extra computational effort is rewarded with a smaller statistical uncertainty. It is therefore interesting to compare the achieved uncertainty with the computational effort, which we may measure as the total number of likelihood samples ($N_{\rm tot} = N_{\rm nest} + M$). Figure 4 shows this comparison for different numbers of live points, given a fixed stopping threshold $\epsilon = 0.01$. The fractional uncertainty clearly decreases with the total number of samples as $\sigma_Z/Z \propto N_{\rm tot}^{-1/2}$, just as expected for a statistical sampling procedure (Skilling 2004, 2006; Chopin & Robert 2008).

In the examples presented here, I have used a low ϵ threshold to require that the live evidence be negligible at the end of the run. Figure 3 suggests, however, that ϵ could be set higher provided that the live evidence is accounted for properly (using the methods in § 2.4).

The lessons here are familiar from previous work on nested sampling, but worth reiterating. The ultimate statistical uncertainty depends mainly on the number of live points. Once the nested sampling procedure has converged (as measured, for example, by the ϵ threshold), running more steps will not improve the results. The way to reduce the uncertainties is to increase the number of live points.[2] That will increase the number of steps it takes to reach convergence, but will yield uncertainties that scale as $\sigma_Z \propto N_{\rm tot}^{-1/2}$.

3.5 Log-normal test case

Nowhere in the theoretical framework was it necessary to specify the form of the likelihood, so the Gaussian test case should be sufficient to validate the analytic results. Nevertheless, it is useful to consider a different test to verify that the results are indeed robust. I use a multivariate log-normal distribution because it is skewed and non-Gaussian but still analytically tractable. Choosing appropriate scaled coordinates, we can write the likelihood in canonical form,

$$\mathcal{L}(\theta) = \prod_{i=1}^{d} \frac{e^{-(\ln\theta_i)^2/2}}{(2\pi)^{1/2}\theta_i} \tag{38}$$

where d is the number of dimensions. Let the prior volume be the cube with $0 < \theta_i < s$, so $V = s^d$ and the evidence is

$$Z = V^{-1} \prod_{i=1}^{d} \int_0^s \frac{e^{-(\ln\theta_i)^2/2}}{(2\pi)^{1/2}\theta_i}\, d\theta_i \tag{39}$$

[2] It is not necessary to start from scratch in order to increase the number of live points. Skilling (2006) explains that independent runs with M_1, M_2, ... live points can be merged into a joint run that effectively has $M_1 + M_2 + \ldots$ live points. The likelihood samplings are simply merged and sorted, while the volume sampling must be recomputed.

For a large prior box, $VZ \approx 1$ and the information gain is

$$H = \frac{1}{VZ} \int \mathcal{L} \ln \frac{\mathcal{L}}{Z} \, d\theta \approx -\frac{d}{2} \left(1 + \ln 2\pi\right) - \ln Z \tag{40}$$

I again work in $d = 4$ dimensions, but now use a box with $s = 20$ to encompass the bulk of the likelihood. Given the larger volume, I use $M = 600$ lives points and take $N_{\rm nest} = 9000$ steps. For these parameter choices, the information gain is $H = 6.31$ and Skilling's estimator for the fractional uncertainty in the evidence has an analytic value of $\sqrt{H/M} = 0.103$.

I first consider a single likelihood sampling and examine the distribution of evidence values for 1000 volume realisations (qv. § 3.2). The empirical mean and standard deviation over the volume samplings are 0.939 ± 0.097. The analytic mean is 0.944, while Skilling's and my estimators predict uncertainties of 0.097 and 0.098, respectively. The histogram of Z values (not shown) agrees well with a Gaussian distribution whose mean and variance are given by eqs. (17) and (23). (Note that Z can have a nearly-Gaussian distribution even if the likelihood is non-Gaussian.)

I next consider 1000 likelihood samplings and examine the distribution of $\langle VZ \rangle_t$ values (qv. § 3.3). The empirical mean and standard deviation are 0.997 ± 0.106. Skilling's and my estimators predict uncertainties of 0.102 and 0.103, respectively. The histogram of $\langle VZ \rangle_t$ again agrees well with the predicted Gaussian distribution. I conclude that the analytic results are reliable even for a non-Gaussian likelihood distribution.

4 SUMMARY

I have derived simple analytic expressions for the mean and variance of the Bayesian evidence over all realisations of the volume sampling in nested sampling, and compared them with the uncertainty estimator introduced by Skilling (2004, 2006) from an information theoretic argument. The two estimators have different forms as sums over the likelihood sampling, yet they yield very similar quantitative results. At this point it is not clear whether the agreement reflects some general equivalence between the two estimators that is not yet apparent, or whether it somehow depends on statistical properties of the likelihood sampling $\{L_i\}$ that emerges from the nested sampling procedure. The moments-based estimator currently has a more rigorous foundation than the information theoretic estimator, but both are useful and it will be interesting to see if they continue to give similar results as nested sampling is applied to a broader range of problems, and if any formal equivalence can be established. Both estimators can be used to compute the statistical uncertainty in the evidence for any implementation that maintains the core prescription of nested sampling: each new point is drawn from the prior distribution in the region inside the current likelihood surface. With these results, determining not only the mean evidence but also the uncertainty requires no additional computational effort (and no guesswork) beyond that needed to generate the likelihood sampling.

ACKNOWLEDGMENTS

I thank Ross Fadely for valuable discussions about nested sampling, and the referee for suggesting the error analysis in § 2.4. This work received support from the US National Science Foundation through grant AST-0747311.

REFERENCES

Betancourt M. J., 2010, e-print arXiv:1005.0157
Brewer B. J., Pártay L. B., Csányi G., 2009, e-print arXiv:0912.2380
Chopin N., Robert C., 2008, e-print arXiv:0801.3887
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Gelman A., Meng X.-L., 1998, Statistical Science, 13, 163
Gelman A. B., Carlin J. S., Stern H. S., Rubin D. B., 1995, Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton
Kullback S., 1959, Information Theory and Statistics. John Wiley and Sons, NY
Mukherjee P., Parkinson D., Liddle A. R., 2006, ApJ, 638, L51
Shaw J. R., Bridges M., Hobson M. P., 2007, MNRAS, 378, 1365
Skilling J., 2004, in American Institute of Physics Conference Series, Vol. 735, American Institute of Physics Conference Series, R. Fischer, R. Preuss, & U. V. Toussaint, ed., pp. 395–405
—, 2006, Bayesian Analysis, 1, 833
—, 2009, in American Institute of Physics Conference Series, Vol. 1193, American Institute of Physics Conference Series, P. M. Goggans & C.-Y. Chan, ed., pp. 277–291